

Dependent Dirichlet Process Rating Model (DDP-RM)

Ken Akira Fujimoto
and
George Karabatsos
University of Illinois-Chicago

12/20/2012

Acknowledgements: This research is supported by NSF research grant SES-1156372, from the Program in Methodology, Measurement, and Statistics. This paper will be presented at the National Council for Measurement in Education (NCME) Conference, April 26-30, at San Francisco. Also, we thank Professor Stephen G. Walker, University of Kent, for helpful conversations about the MCMC algorithm we used for the paper.

Dependent Dirichlet Process Rating Model (DDP-RM)

Abstract

Rating scale items are ubiquitous in psychometric practice. Yet, the psychometric properties of the rating scales can often vary by examinee, as well as by item. To address this practical psychometric problem, we introduce a novel, Bayesian nonparametric IRT model for rating scale items. The model is an infinite-mixture of Rasch partial credit models, with rating thresholds being the random parameters that are subject to the mixture, and with (infinitely-many) covariate-dependent stick-breaking weights. Random parameters and the mixture weights are assigned a Dependent Dirichlet process prior (DDP) distribution. Thus, the novel model allows the rating category thresholds to vary flexibly across items and examinees, and allows the distribution of the category thresholds to vary flexibly as a function of covariates. We illustrate the novel model through the analysis of a real rating data set that has been studied extensively in the psychometric modeling literature. The model is shown to have better predictive performance than other IRT rating models of common usage.

KEYWORDS: Rating Scale Analysis, Bayesian Nonparametrics, Bayesian Inference

RUNNING TITLE: Dependent Dirichlet Process Rating Model.

Acknowledgements: This research is supported by NSF research grant SES-XXXXXXX, from the Program in Methodology, Measurement, and Statistics.

1 Introduction

Item response theory (IRT) provide an modeling approach for estimating the psychometric properties of data gathered with tests consisting of rating scale items. An IRT rating scale model provides useful information about the qualities of the given test, such as the difficulty of each test item, and the response thresholds of the rating categories. Typical IRT rating models of common usage include the Rasch rating scale model (Andrich, 1978), partial credit models (Masters, 1982; Muraki, 1992), and the family of graded response models (Samejima, 1969, 1972). All of these IRT models have seen many successful applications for a wide range of research settings.

Still, there is room to further develop IRT rating models. Typical IRT rating models assume that the rating category threshold parameters do not vary across examinees. However, the assumption is questionable, as it is reasonable to view that rating scale usage also varies across examinees. In summary, if the assumption is empirically violated, then the IRT rating model may poorly fit the data. Moreover, in practice, it may also be of interest to investigate how rating scale usage (thresholds) vary across examinees, in order to investigate for differential item functioning (DIF).

In principle, one may relax this assumption through the specification of a discrete-mixture IRT model, with the rating threshold parameters being the random parameters that are subject to the mixture. In general, a discrete mixture model has the form (e.g., McLachlan & Peel, 2000):

$$f_{G_{\mathbf{x}}}(y|\mathbf{x}) = \int f(y|\mathbf{x}; \Psi(\mathbf{x}))dG_{\mathbf{x}}(\Psi) = \sum_{h=1}^H f(y|\mathbf{x}; \Psi_h(\mathbf{x}))\omega_h(\mathbf{x}),$$

given, possibly covariate (\mathbf{x}) dependent, mixing distribution $G_{\mathbf{x}}$, component indices $h = 1, \dots, H$, kernel (component) densities $f(y|\mathbf{x}; \Psi_h(\mathbf{x}))$ ($h = 1, \dots, H$), and mixing weights $(\omega_h(\mathbf{x}))_{h=1}^H$ which sum to 1 at every $\mathbf{x} \in \mathcal{X}$. A mixture IRT rating model treats $y \in \{k = 0, 1, \dots, m\}$ as a rating, and each of the kernel densities $f(y|\mathbf{x}; \Psi_h(\mathbf{x})) = f(y; \theta, \tau_h)$ ($h = 1, \dots, H$) are chosen as an IRT model, such as the Rasch partial credit model (PCM):

$$f(y|\theta, \tau_h) = P(Y = y|\theta, \tau_h) = \frac{\exp(y\theta - \sum_{l=0}^y \tau_{lh})}{\sum_{k=0}^m \exp(k\theta - \sum_{l=0}^k \tau_{lh})},$$

where $(\tau_{0h}, \dots, \tau_{mh})$ are the rating category threshold parameters corresponding to the h th mixture component. Also, in typical IRT mixture models, the examinee ability parameter θ is assigned a normal prior distribution. However, none of the available mixture IRT rating models model the rating category thresholds as parameters that are subject to the mixture (Rost, 1991; Smit, Kelderman, & van der Flier, 2003; Von Davier & Yamamoto, 2004; Frick, Strobl, Leisch, & Zeileis, 2012). Moreover, most of these models are finite mixtures (i.e., $H < \infty$), which have limited flexibility to adequately describe many rating scale data sets. We could achieve maximum modeling flexibility in a fully nonparametric framework, through the specification of an infinite-mixture model (i.e., $H = \infty$).

To address all these practical limitations of existing IRT models, we introduce a novel, Bayesian nonparametric IRT rating model. The model is an infinite-mixture of Rasch par-

tial credit models, with rating category threshold parameters subject to the mixture, and with infinitely-many covariate-dependent stick-breaking weights. Specifically, the random parameters and the mixture weights are modeled by a Dependent Dirichlet process (DDP) (MacEachern, 1999; 2000; 2001), namely the local Dirichlet process (LDP) (Chung & Dunson, 2011). Therefore, we refer to our novel model as the DDP Rating Model (DDP-RM).

Our model adds to the body of literature that incorporate Dirichlet process (DP) priors for IRT, such as DP mixtures of 3-parameter logistic model item parameters (Miyazaki & Hoshino, 2009), DP mixtures of Rasch model ability parameters (San Martin et al., 2011), and a DDP model for the link function of the 2-parameter IRT model (Duncan & MacEachern, 2008). However, all of these models focus on dichotomous item scores, and nearly all of these models assume a less-flexible DP, which assumes no covariate-dependence. Karabatsos and Walker (2012 to appear) provides a review of DP and DDP mixture models for IRT.

In Section 2, we introduce our Bayesian nonparametric infinite-mixture IRT model for rating scale data, after giving a necessary review of the DP, the DDP, and the general DDP infinite-mixture model. The Appendix A describes the Markov Chain Monte Carlo (MCMC) algorithm that is used to estimate the posterior distribution of our model. In Section 3, we illustrate our model on a real data set of rating scale items, which has been extensively studied in the psychometric modeling literature (De Boeck & Wilson, 2004). There, we also compare the goodness of predictive fit of our model against other IRT rating scale models of common usage. Section 4 ends with conclusions, and with some discussion about future modeling extensions.

2 Nonparametric Infinite-mixture IRT

2.1 Dependent Dirichlet Process (DDP)

Throughout, $n_p(\cdot|\cdot, \cdot)$, $ga(\cdot|\cdot, \cdot)$, $beta(\cdot|\cdot, \cdot)$, and $un(\cdot|\cdot, \cdot)$ denote the density functions of the p -variate Normal $_p(\cdot|\cdot, \cdot)$, Gamma $(\cdot|\cdot, \cdot)$, Beta $(\cdot|\cdot, \cdot)$, and Uniform $(\cdot|\cdot, \cdot)$ distribution functions, respectively, where the gamma distribution is parameterized by shape and scale parameters.

We first review the DP, which is the basis of the DDP. Let G denote a random distribution. If this random distribution is a Dirichlet process (DP), it is denoted by $G \sim DP(\alpha, G_0)$, with precision parameter $\alpha > 0$, with baseline distribution G_0 , mean $E[G(\cdot)] = G_0(\cdot)$, and variance $Var[G(\cdot)] = G_0(\cdot)[1 - G_0(\cdot)]/(\alpha + 1)$ (Ferguson, 1973). Also, a Dirichlet process $G \sim DP(\alpha, G_0)$ random distribution can be constructed according to a stick-breaking process (Sethuraman, 1994), via:

$$G(\cdot) = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_h}(\cdot), \quad (1)$$

according to stick-breaking weights defined by:

$$\omega_h = v_h \prod_{r=1}^{h-1} (1 - v_r), \quad h = 1, 2, \dots$$

which sum to 1 almost surely, given random variates

$$v_1, v_2, \dots \sim_{i.i.d.} \text{Beta}(1, \alpha); \quad \theta_1, \theta_2, \dots \sim_{i.i.d.} G_0,$$

where $\delta_{\theta_h}(\cdot)$ denotes a point-mass distribution function that assigns probability 1 to the value θ_h . Hence, a random DP distribution, $G \sim \text{DP}(\alpha, G_0)$, is formed by an infinite-mixture of such point-mass distributions, and is almost surely discrete.

The DDP extends the DP to regression settings, by providing a model for a covariate-dependent random distribution $G_{\mathbf{x}}$ (MacEachern, 1999, 2000, 2001). A DDP has the stick-breaking representation:

$$G_{\mathbf{x}}(\cdot) = \sum_{h=1}^{\infty} \omega_h(\mathbf{x}) \delta_{\theta_h(\mathbf{x})}(\cdot), \quad (2)$$

where the stick-breaking weights $\omega_h(\mathbf{x}) = v_h(\mathbf{x}) \prod_{r=1}^{h-1} (1 - v_r(\mathbf{x}))$ ($h = 1, 2, \dots$) sum to 1 at every covariate value $\mathbf{x} \in \mathcal{X}$, with $v_r(\mathbf{x}) : \mathcal{X} \rightarrow [0, 1]$. The specification of a DDP model is completed by the specification of a prior distribution for the mixture weights $\{\omega_h(\mathbf{x})\}_{h=1,2,\dots}$ and atoms $\{\theta_h(\mathbf{x})\}_{h=1,2,\dots}$, which are infinite collections of processes indexed by the \mathbf{x} -space. Such a prior distribution is called a DDP prior, and is a prior for the distribution $G_{\mathbf{x}}$.

An example of a DDP prior is the local DP (lDP; Chung & Dunson, 2011). The lDP makes use of a predictor-dependent set $\mathcal{L}_{\mathbf{x}} = \{h : d(\mathbf{x}, \Gamma_h) < \psi\} \subset \{1, 2, \dots\}$, which indexes the locations belonging to the neighborhood \mathbf{x} of size $\psi > 0$, and $d(\cdot, \cdot)$ is a chosen distance measure (e.g., Euclidean). Then the lDP models stick-breaking mixture weights $\{\omega_h(\mathbf{x})\}_{h=1,2,\dots}$ and atoms $\{\theta_h(\mathbf{x})\}_{h=1,2,\dots}$ as locally covariate-dependent, through the specification of local random components

$$\Gamma(\mathbf{x}) = \{\Gamma_h, h \in \mathcal{L}_{\mathbf{x}}\}, \quad \mathbf{v}(\mathbf{x}) = \{v_h, h \in \mathcal{L}_{\mathbf{x}}\}, \quad \boldsymbol{\theta}(\mathbf{x}) = \{\theta_h, h \in \mathcal{L}_{\mathbf{x}}\}.$$

Specifically, the lDP constructs a random covariate-dependent mixing distribution:

$$G_{\mathbf{x}}(\cdot) = \sum_{h \in \mathcal{L}_{\mathbf{x}}} \omega_h(\mathbf{x}) \delta_{\theta_h(\mathbf{x})}(\cdot),$$

on the basis of infinitely-many mixture weights that are defined by:

$$\omega_h(\mathbf{x}) = v_{\pi_h(\mathbf{x})} \prod_{\{l \in \mathcal{L}_{\mathbf{x}} : \pi_l(\mathbf{x}) < \pi_h(\mathbf{x})\}} (1 - v_{\pi_l(\mathbf{x})}),$$

where $|\mathcal{L}_{\mathbf{x}}|$ is the cardinality of the set $\mathcal{L}_{\mathbf{x}}$, and the $\{\pi_l(\mathbf{x}) : l \in \mathcal{L}_{\mathbf{x}}\}$ indicate the ordering of the indices $l \in \mathcal{L}_{\mathbf{x}}$ on the basis of \mathbf{x} . The lDP is completed by the specification of a prior distribution on $\{\Gamma(\mathbf{x}), \mathbf{v}(\mathbf{x}), \boldsymbol{\theta}(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, to define a DDP prior on $G_{\mathbf{x}}$.

Using a DDP, it is possible to construct an infinite-mixture model via

$$f(y|\mathbf{x}) = \int f(y|\mathbf{x}; \boldsymbol{\Psi}(\mathbf{x})) dG_{\mathbf{x}}(\boldsymbol{\Psi}) = \sum_{h=1}^{\infty} f(y|\mathbf{x}; \boldsymbol{\Psi}_h(\mathbf{x})) \omega_h(\mathbf{x}), \quad (3)$$

where the $f(y|\mathbf{x}; \boldsymbol{\Psi}_h(\mathbf{x}))$ represent chosen kernel densities, and where the covariate-dependent

random mixing distribution $G_{\mathbf{x}}$ is assigned a DDP prior, and the $\omega_h(\mathbf{x})$ are covariate-dependent stick-breaking weights. Such a model is referred to as a DDP mixture model, extending the idea of a DP mixture model (Lo, 1984) which assumes that the mixture distribution G is not covariate-dependent.

2.2 The Dependent Dirichlet Process Rating Model (DDP-RM)

The DDP-RM is a DDP mixture model defined by:

$$P(Y = y|\mathbf{x};\theta, \gamma) = \int f(y|\theta, \tau) dG_{\mathbf{x}}(\tau) = \sum_{h=1}^{\infty} f(y|\theta_h, \tau_h) \omega_h(\mathbf{x}^\top \gamma), \quad (4)$$

where the kernels $f(y|\theta, \tau_h)$ are specified by the Rasch partial credit model:

$$f(y|\theta, \tau_h) = P(Y = y|\theta, \tau_h) = \frac{\exp(y\theta - \sum_{l=0}^y \tau_{lh})}{\sum_{k=0}^m \exp(k\theta - \sum_{l=0}^k \tau_{lh})},$$

where for the $m+1$ rating categories $k = 0, 1, \dots, m$, we assume the constraint $\tau_{0h} \equiv 0$, with free threshold parameters $\tau_h = (\tau_{1h}, \dots, \tau_{mh})$, for the h^{th} mixture component. Also, throughout, we denote the ability parameter of a given examinee t is by θ_t . Also, the stick-breaking mixture weights $\{\omega_h(\mathbf{x})\}_{h=1,2,\dots}$ and atoms $\{\tau_h(\mathbf{x})\}_{h=1,2,\dots}$ are modeled by an IDP prior; more details later. Moreover, \mathbf{x} can be a general vector of covariates, which may for example, either describe examinee characteristics (gender, race, and/or social economic status), or describe test characteristics (time at which item was administered, item type, etc.). For the next section, where we provide an empirical illustration of our model, we consider the case where \mathbf{x} are item (0-1) indicators.

The mixing distribution in our model is formed according to the following novel modification of the IDP (Chung & Dunson, 2011). First let

$$\mathcal{L}_{\mathbf{x}} = \{h : d(\mathbf{x}^\top \gamma, h) < \psi(\mathbf{x})\} \subset \{1, 2, \dots\}$$

denote the subset of mixture component indices $h \in \mathbb{Z}^+$ having fixed addresses $\{\Gamma_h = h\}$ which are within a $\psi(\mathbf{x})$ -neighborhood around the linear predictor $\mathbf{x}^\top \gamma$. Then under our formulation of the IDP, the local variables are defined by $\mathbf{v}(\mathbf{x}^\top \gamma) = \{v_h, h \in \mathcal{L}_{\mathbf{x}}\}$ for the specification of stick-breaking mixture weights

$$\omega_h(\mathbf{x}^\top \gamma) = v_h(\mathbf{x}^\top \gamma) \prod_{\{l \in \mathcal{L}_{\mathbf{x}} : l < h\}} (1 - v_l(\mathbf{x}^\top \gamma)), \quad h \in \mathcal{L}_{\mathbf{x}}, \quad (5)$$

and defined by rating category threshold atoms $\tau(\mathbf{x}^\top \gamma) = \{\tau_h, h \in \mathcal{L}_{\mathbf{x}}\}$. We fix $v_{\max(\mathcal{L}_{\mathbf{x}})}(\mathbf{x}^\top \gamma) \equiv 1$ to ensure that the mixture weights $\omega_h(\mathbf{x}^\top \gamma)$ sum to 1 for each \mathbf{x} (Chung & Dunson, 2011). Thus, our IDP forms stick-breaking mixture weights, by selecting the strict subset of stick-breaking parameters ($\{v_h\}$) and atoms ($\{\tau_h\}$) that are within neighborhood centered around (a linearized) \mathbf{x} . For example, when $\mathbf{x}^\top \gamma = 10$ and $\psi(\mathbf{x}) = 2.5$, then the covariate (\mathbf{x})-dependent local subset becomes $\mathcal{L}_{\mathbf{x}} = \{8, 9, 10, 11, 12\}$.

Then the mixture weights of equation (5) give rise to a covariate-dependent mixing dis-

tribution:

$$G_{\mathbf{x}}(\cdot) = \sum_{h \in \mathcal{L}_{\mathbf{x}}} \omega_h(\mathbf{x}^\top \boldsymbol{\gamma}) \delta_{\boldsymbol{\tau}_h(\mathbf{x}^\top \boldsymbol{\gamma})}(\cdot).$$

Thus, for two covariates \mathbf{x} and \mathbf{x}' , the level of similarity between $\mathcal{L}_{\mathbf{x}}$ and $\mathcal{L}_{\mathbf{x}'}$ determines the level of similarity between the two corresponding mixing distribution $G_{\mathbf{x}}(\cdot)$ and $G_{\mathbf{x}'}(\cdot)$, with the level of similarity controlled by the parameters $(\boldsymbol{\gamma}, \boldsymbol{\psi}(\mathbf{x}))$.

The DDP-RM is completed by the specification of the following prior distributions:

$$\begin{aligned} \theta_t &\sim \text{Normal}(0, 1), \\ \boldsymbol{\tau}_h &\sim \text{Normal}_{m_j-1}(\mathbf{0}, \boldsymbol{\Sigma}_\tau), \\ v_h &\sim \text{Beta}(1, \alpha), \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha), \\ \boldsymbol{\gamma} &\sim \pi_\gamma, \\ \boldsymbol{\psi}(\mathbf{x}) &\sim \pi_{\mathbf{x}}. \end{aligned}$$

where π_γ and $\pi_{\mathbf{x}}$ are generic prior densities. In the next section, where we illustrate our model, we consider a useful choice of the prior distributions listed above.

The unique features of our model is that it clusters item category thresholds based on similar mixing distribution, which is captured through the neighborhood inducing parameter $\boldsymbol{\gamma}$. When two separate $\boldsymbol{\gamma}$ s have the same values, then the mixture components are the same for the covariates associates with the two $\boldsymbol{\gamma}$ s. In the present study, because the covariates are item indicators, similar $\boldsymbol{\gamma}$ s would suggest that the items associated with the $\boldsymbol{\gamma}$ s have similar mixing distributions describing the random relative category thresholds, thus possibly suggesting that a common set of thresholds could be specified for this group of items. Another unique feature of our model is that it forms the mixing distribution nonparametrically and allows the mixing distribution to depend on covariates.

For notational convenience, denote a sample set of rating data by $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n=NJ}$, provided by N examinees ($t = 1, \dots, N$) on J test items ($j = 1, \dots, J$), and with $n = NJ$ giving the total number of item responses in the data set. Each $y_i \in \mathcal{D}_n$ denotes a rating by a particular examinee on a particular item. According to standard arguments of probability theory involving Bayes' theorem, given a data set \mathcal{D}_n having likelihood $\prod_{i=1}^n f(y_i|\mathbf{x}_i; \boldsymbol{\zeta})$ under our model with parameters $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{v}, \alpha, \boldsymbol{\gamma}, \boldsymbol{\psi})$, and given a proper prior density $\pi(\boldsymbol{\zeta})$ defined over the space $\Omega_{\boldsymbol{\zeta}}$ of $\boldsymbol{\zeta}$, the posterior density of $\boldsymbol{\zeta}$ is proper and given by:

$$\pi(\boldsymbol{\zeta}|\mathcal{D}_n) \propto \prod_{i=1}^n P(y_i|\mathbf{x}_i; \boldsymbol{\zeta})\pi(\boldsymbol{\zeta})$$

up to a proportionality constant. Then the posterior predictive density of Y for a chosen \mathbf{x} is given by:

$$f_n(y|\mathbf{x}) = \int f(y|\mathbf{x}; \boldsymbol{\zeta})\pi(\boldsymbol{\zeta}|\mathcal{D}_n)d\boldsymbol{\zeta},$$

with this density corresponding to posterior predictive mean (expectation) and variance (Var)

$$E_n(Y|\mathbf{x}) = \int y f_n(y|\mathbf{x})dy, \quad \text{Var}_n(Y|\mathbf{x}) = \int \{y - E(Y|\mathbf{x})\}^2 f_n(y|\mathbf{x})dy.$$

We make use of the MCMC sampling methods for Bayesian infinite-mixture models that are described by Kalli, Griffin, and Walker (2011), to perform inference of the posterior density $\pi(\boldsymbol{\zeta}|\mathcal{D}_n)$ and posterior predictive density $f_n(y|\mathbf{x})$ of the model, and inference of all posterior functionals of interest. These methods are based on the use of strategic latent variables. As mentioned, the Appendix A provides more details about all the conditional posterior distribution of the model, which are sampled at each stage of the MCMC algorithm.

2.3 Model Assessment of Predictive Performance

Given a set of data \mathcal{D}_n , suppose it is of interest to compare the predictive performance between \underline{M} different IRT rating models, with each model indexed by $\underline{m} = 1, \dots, \underline{M}$. It is possible to compare the predictive performance between the models using a mean-squared predictive error criterion, namely the $D_1(\underline{m})$ criterion (Gelfand & Ghosh, 1998). For a given model $\underline{m} \in \{1, \dots, \underline{M}\}$ under comparison, the criterion is defined by:

$$\begin{aligned} D_1(\underline{m}) &= \sum_{i=1}^n [y_i - E_n(Y_i|\mathbf{x}_i, \underline{m})]^2 + \sum_{i=1}^n \text{Var}_n(Y_i|\mathbf{x}_i, \underline{m}) \\ &= \text{GF}(\underline{m}) + \text{Pen}(\underline{m}) \end{aligned}$$

In the equation above, the first term is a predictive bias measure which indicates the goodness-of-fit ($\text{GF}(\underline{m})$) of the model. The second term is a penalty which is large when the model is either over-fitting or under-fitting the given data set \mathcal{D}_n . For all other comparison models, the $E_n(Y_i|\mathbf{x}_i, \underline{m})$ and $\text{Var}_n(Y_i|\mathbf{x}_i, \underline{m})$ are derived from marginal maximum or conditional maximum likelihood parameter estimates. For a non-Bayesian model having point estimate $\hat{\boldsymbol{\zeta}}_n = \hat{\boldsymbol{\zeta}}(\mathcal{D}_n)$, such as a maximum-likelihood estimation, the criterion is estimated via $\hat{E}_n(Y_i|\mathbf{x}_i, \underline{m}) = E(Y_i|\mathbf{x}_i, \underline{m}, \hat{\boldsymbol{\zeta}}_n)$ and $\hat{\text{Var}}_n(Y_i|\mathbf{x}_i, \underline{m}) = \text{Var}(Y_i|\mathbf{x}_i, \underline{m}, \hat{\boldsymbol{\zeta}}_n)$ ($i = 1, \dots, n$) (Gelfand & Ghosh, 1998).

3 Model Illustration

In this section, we compared the predictive performance of the DDP-RM to several other IRT rating models, on a real data set obtained from the verbal aggression study (De Boeck & Wilson, 2004). The verbal aggression data set contains the item ratings 316 students (243 females and 73 males) from a Dutch-speaking Belgian university. Each student rated 24 items, which are indicators for levels of verbal aggression (e.g., “A bus fails to stop for me. I would want to curse.”), on a scale of 0 = no, 1 = perhaps, and 2 = yes. The items are categorized into a $2 \times 2 \times 3$ design: Behavior Mode (Want or Do) by Situation Type (Other-to-blame or Self-to-blame) by Behavior Type (Curse, Scold, or Shout). Appendix B lists all 24 items.

3.1 Model Specifications and MCMC Diagnostics

When applying the DDP-RM to analyze the verbal aggression data set, we assigned a proper prior $\theta_t \sim_{iid} \text{n}(0, 1)$, along with high-variance proper priors $\boldsymbol{\tau}_h \sim_{iid} \text{n}(\mathbf{0}, 5\mathbf{I}_m)$, $\boldsymbol{\gamma} \sim \text{un}(1, 745)$,

$\psi(\mathbf{x}_i) \sim_{iid} \text{un}(.5, 20)$, to reflect the limited prior information about these model parameters.

We ran the MCMC sampling algorithm for 200,000 MCMC sampling iterations. We discarded the first 100,000 MCMC samples (i.e., burn-in period), and saved every fifth sample thereafter, saving a total of 20,000 posterior samples. We used standard methods to examine whether our MCMC algorithm (presented in the Appendix A) generated a sufficiently-large number of samples from the posterior distribution. Given a finite S number of samples $\left\{\zeta^{(s)}\right\}_{s=1}^S$ generated by the MCMC algorithm, univariate trace plots can be used to evaluate the mixing of the chain (i.e. the extent to which the chain explores the support of the posterior distribution). Also, batch means methods can be used (Jones, 2006) to estimate the 95% MC confidence interval (MCCI), for estimates of marginal posterior moments.

In Figures 1 and 2, we present the trace plots of the MCMC samples of the threshold estimates for three items and ability estimates for six examinees. The trace plots suggest that the estimates for all parameters stabilized after the burn-in period. The trace plots for all other parameter estimates were similar. The posterior means for all parameter also had sufficiently small 95% Monte Carlo (MC) confidence intervals according to a consistent batch means estimator. For each item, the 95% MC confidence interval half-width for the posterior means and standard deviations for the category threshold estimates are presented in Table 1.

The category thresholds estimates ranged from -0.68 to 3.32 . Similar to the conclusions in De Boeck and Wilson (2004), Item 21 was the most difficult (i.e., the largest estimates for the category thresholds), which suggest that examinees require a higher level of verbal aggression to endorse the higher rating categories for this item. Item 4 was the easiest (i.e., the smallest estimates for the category thresholds), which suggest that examinees require lower levels of verbal aggression to endorse the higher rating categories for this item.

A unique feature of the DDP-RM is the information it provides in the posterior predictive density about how examinees used the rating categories. That is, were different category threshold levels of clusters of examinees present for an item? We display in Figure 3 for three items. Notice that Items 1 and 23 exhibit greater variability in their densities compared to Item 2. Moreover, the density for Item 1’s first threshold has a tri-modal form (i.e., one major and two minor), suggesting three clusters of examinees with different levels of category thresholds are present. The density for the second threshold for this item is bimodal. The densities for Item 23’s first and second thresholds have a second minor mode, though the second mode is larger for the second threshold. The densities for Item 2, on the other hand, have only a single mode, suggesting that a single cluster of examinees exists with respect to category threshold estimates. Moreover, notice that the variability for this item is much smaller. Thus, a single set of category threshold estimates is much more appropriate to represent all examinees compared to the other Items 1 and 23. Most of the item’s posterior predictive densities consisted of a single mode. For the values of the model for each threshold by item, please refer to Table 1. Traditional models do not provide such information. With traditional models, one only has the threshold estimates to compare across items, which could lead to misleading conclusions, as might be the case in this situation.

Through the neighborhood location and size parameters (i.e., γ and ψ , respectively), the DDP-RM also provides information about the similarities in mixing distributions across items. Neighborhood location estimates ranged from 6.0 to 255.6. The neighborhood size es-

estimates ranged from 7.5 to 19.8. None of the items have the same neighborhood location and size estimates (i.e., $\gamma_j \neq \gamma_{j'}$ and $\psi_j \neq \psi_{j'}$ for two different j s). Thus, none of the items share a common mixing distribution. For all items, the 95% MC confidence interval half-width ranged from .02 to .93 for the posterior mean and .01 to .79 for the posterior standard deviation for the neighborhood location γ . The 95% MC confidence interval half-width ranged from .01 to .93 for the posterior mean and .01 to .79 for the posterior standard deviation for the neighborhood size ψ . For the median and quartile range for the neighborhood location and size, please refer to Figure 4.

As with other common IRT models for rating data, the DDP-RM provides posterior means of examinee abilities, which represent the examinees' level on the latent trait scale. The examinee ability estimates ranged from -2.37 to 3.74 , with a mean of -0.02 and standard deviation of 1.01 . The 95% MC confidence interval half-width ranged from .01 to .03 for the mean of the posterior ability estimates, and .00 to .03 for the standard deviation of the posterior ability estimates.

3.2 Model Comparisons

In this study, the comparison models were the partial credit model (PCM) (Masters, 1982), generalized partial credit model (GPCM) (Muraki, 1992), rating scale model (RSM) (Andrich, 1978), graded response model (GRM) (Samejima, 1969), nominal response model (NRM) (Bock, 1972), mix partial credit model (mix-PCM) (Rost, 1991), and a covariate-independent DP mixture PCM model that treated the category thresholds as random. All models except the latter two were fit using IRTPRO 2.1 (Cai, Thissen, & du Toit, 2011). The mix-PCM was fit in WINMIRA 2001 (von Davier, 2001), and a 3-mixture PCM provided a best fit according to the AIC model selection criterion (Akaike, 1973). Based on preliminary analyses, a three-mixture PCM achieved the best predictive performance compared to a one-, two-, four-, and five-mixture PCM. Thus, we report the predictive performance of the three-mixture PCM. The DP mixture PCM model was fit in MATLAB (2012, The MathWorks, Natick, MA). The baseline distribution for the set of m thresholds was distributed as a multivariate normal distribution with density $n(\mathbf{0}, \mathbf{I}_m)$; the examinee abilities were assumed to follow a univariate normal with density $n(0, 1)$; and the precision parameter α was fixed to 1.

For the DDP-RM and the DP mixture PCM, the 95% Monte Carlo Confidence interval half width were generally less than 1. Moreover, for each $D_1(\underline{m})$, $GF(\underline{m})$, and $Pen(\underline{m})$, there was no overlap between two models, after accounting for the 95% MCCI. Table 2 contains the $D_1(\underline{m})$ for all models included in the analysis of the verbal aggression data set. The DDP-RM outperformed all comparison models by at least 49 $D_1(\underline{m})$ units. In all, the three mixture models outperformed the traditional, single-mixture models, which suggests that more than one latent class is present in the data set. The finite-mixture Rasch PCM model, while outperformed the single-mixture models, was still is bested by the two infinite-mixture models. The DDP-RM outperforming the DP-mixture PCM suggests that all items do not share a common mixing distribution.

4 Conclusions

We have introduced a novel Bayesian nonparametric rating scale IRT model, named the DDP-RM, which is an infinite-mixture model that is based on the local Dirichlet process formulation of the DDP. The model, through the posterior predictive distribution, describes how the examinees are using the rating categories. Specifically, it can reveal the number of possible groups of examinees that may be present for a given item threshold based on the number of modes displayed in the distribution. Moreover, we demonstrated that the new model can provide a substantially-better predictive fit of the rating data, compared to other IRT models of common usage.

In future research, it would be of interest to extend the DDP-RM, to have (infinitely-many) mixture weights that are more flexible than the stick-breaking weights of the DDP. For example, Karabatsos and Walker (2012) proposed novel mixture weights that are based on an infinite-ordered probits regression model with covariate dependence in the mean and variance. Alternatively, the (infinitely-many) mixture weights can be specified by a covariate-dependent version of normalized random measures (Regazzini, Lijoi & Prünster, 2003; Lijoi, Meña, & Prünster, 2005, 2007; James, Lijoi, & Prünster, 2009).

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, 2nd Tsahkadsor*, Armenian SSR (pp. 267-281).
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Cai, L., du Toit, S., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Chicago, IL: Scientific Software International.
- Chung, Y. & Dunson, D. (2011). The local Dirichlet process. *Annals of the Institute of Statistical Mathematics*, 63, 59-80.
- De Boeck, P. & Wilson, M. (2004). Explanatory item response models: A generalized linear and nonlinear approach. Springer.
- Duncan, K. & MacEachern, S. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, 8, 41-66.
- Escobar, M. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 577-588.
- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209-230.
- Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch mixture models with package psychomix. *Journal of Statistical Software*, 48, 1-25.
- Gelfand, A. & Ghosh, S. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85, 1-11.
- James, L., Lijoi, A., & Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36, 76-97.
- Jones, G., Haran, M., Caffo, B., & Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101, 1537-1547.
- Kalli, M., Griffin, J., & Walker, S. (2011). Slice sampling mixture models. *Statistics and Computing*, 21, 93-105.
- Karabatsos, G. & Walker, S. (2012). Adaptive-modal Bayesian nonparametric regression. *Electronic Journal of Statistics*, 6, 2038-2068.

- Karabatsos, G. & Walker, S. (2012 to appear). Bayesian nonparametric IRT. In W. van der Linden & R. Hambleton (Eds.), *Handbook of Item Response Theory: Models, Statistical Tools, and Applications*. New York: Taylor and Francis.
- Lijoi, A., Mena, R., & Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, *100*, 1278-1291.
- Lijoi, A., Mena, R., & Prünster, I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society, Series B*, *69*, 715-740.
- Lo, A. (1984). On a class of Bayesian nonparametric estimates. *Annals of Statistics*, *12*, 351-357.
- MacEachern, S. (1999). Dependent nonparametric processes. *Proceedings of the Bayesian Statistical Sciences Section of the American Statistical Association* (pp. 50-55).
- MacEachern, S. (2000). *Dependent Dirichlet Processes*. Technical report, Department of Statistics, The Ohio State University.
- MacEachern, S. (2001). Decision theoretic aspects of dependent nonparametric processes. In E. George (Ed.), *Bayesian Methods with Applications to Science, Policy and Official Statistics* (pp. 551-560). Creta: International Society for Bayesian Analysis.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.
- McLachlan, G. & Peel, D. (2000). *Finite mixture models*. Wiley-Interscience.
- Miyazaki, K. & Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika*, *74*, 375-393.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.
- Regazzini, E., Lijoi, A., & Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics*, *31*, 560-585.
- Roberts, G. & Rosenthal, J. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, *18*, 349-367.
- Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *Journal of Mathematical and Statistical Psychology*, *44*, 75-92.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*.

- Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph*, 18.
- San Martín, E., Jara, A., Rolin, J., & Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika*, 76, 385-409.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.
- Smit, A., Kelderman, H., & van der Flier, H. (2003). Latent trait latent class analysis of an Eysenck Personality Questionnaire. *Methods of Psychological Research Online*, 8, 23-50.
- von Davier, M. (2001). *WINMIRA 2001*. [Computer software]. St. Paul, MN: Assessment Systems Corporation.
- von Davier, M. & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, 28, 389-406.

APPENDIX A: MCMC Sampling Methods

We implement the MCMC sampling method of Kalli et al., (2011) to estimate our infinite-mixture IRT model. This MCMC sampling method involves introducing strategic latent variables in order to implement exact MCMC algorithms for the estimation of the model's posterior distribution. That is, for our DDP-RM (Section 2), we introduce the latent variables $(u_i, z_i \in \mathbb{Z})_{i=1}^n$ and a decreasing function $\xi_h = \exp(-h)$, so that the model's data likelihood can be written as the joint distribution:

$$\prod_{i=1}^n f(u_i, z_i, y_i | \mathbf{x}; \boldsymbol{\zeta}) = \prod_{i=1}^n \left\{ \mathbb{I}(0 < u_i < \xi_{z_i}) \xi_{z_i}^{-1} f(y_i | \theta_{t(i)}, \boldsymbol{\tau}_{z_i}) \omega_{z_i}(\mathbf{x}'\boldsymbol{\gamma}) \right\}, \quad (6)$$

where $\theta_{t(i)}$ denotes the ability of examinee t who corresponds to rating y_i , and where $\mathbb{I}(\cdot)$ is the indicator function. Marginalizing over each of the latent variables (u_i, z_i) in Equation 6, for each $i = 1, \dots, n$, returns the original likelihood,

$$\prod_{i=1}^n \left\{ \sum_{h=1}^{\infty} f(y_i | \theta_{t(i)}, \boldsymbol{\tau}_h) \omega_h(\mathbf{x}'\boldsymbol{\gamma}) \right\},$$

of our infinite-dimensional IRT model. Thus, provided the latent variables, the model can be characterized as a finite-dimensional model, which in turn, permits the use of standard MCMC methods to sample the model's full joint posterior distribution. Given all variables, save the $(z_i)_{i=1}^n$, the choice of each z_i has minimum 1 and maximum N_{\max} , where $N_{\max} = \max_i [\max_h \mathbb{I}(u_i < \xi_h) h]$.

Specifically, for each $i = 1, \dots, n$ and $t = 1, \dots, T$, each of the model parameters is sampled from its corresponding full conditional posterior distribution, at each stage s ($s = 1, \dots, S$) of the MCMC algorithm. We assume the prior form as in the empirical illustration of our model, in the analysis of the verbal aggression data set, as in Section 3. The full conditional posterior distribution for each block of model parameters are as follows:

1. $\pi(u_i | \dots) = \text{un}(u_i | 0, \xi_{z_i})$;
2. $\pi(z_i = h | \dots) \propto \mathbb{I}(u_i < \xi_h) \xi_h^{-1} f(y_i | \theta_{t(i)}, \boldsymbol{\tau}_h) \omega_h(\mathbf{x}'\boldsymbol{\gamma})$, $h = 1, \dots, N_{\max}$;
3. $\pi(\theta_t | \dots) \propto \text{n}(\theta | 0, 1) \prod_{i:t(i)=t} f(y_i | \theta_{t(i)}, \boldsymbol{\tau}_{z_i})$;
4. $\pi(\boldsymbol{\gamma} | \dots) \propto \text{un}_p(\boldsymbol{\gamma} | a_{\boldsymbol{\gamma}}, b_{\boldsymbol{\gamma}}) \prod_{i=1}^n v_{z_i} \prod_{l=1}^{z_i-1} (1 - v_l)$;
5. $\pi(\boldsymbol{\psi}(\mathbf{x}) | \dots) \propto \text{un}(\boldsymbol{\psi}(\mathbf{x}) | a_{\boldsymbol{\psi}}, b_{\boldsymbol{\psi}}) \prod_{i=1}^n v_{z_i} \prod_{l=1}^{z_i-1} (1 - v_l)$;
6. $\pi(\boldsymbol{\tau}_h | \dots) \propto \text{n}_m(\boldsymbol{\tau}_h | \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\tau}}) \prod_{i \in h} f(y_i | \theta_{t(i)}, \boldsymbol{\tau}_{z_i})$, $h = 1, \dots, N_{\max}$;
7. $\pi(v_h | \dots) = \text{beta}\left(v_h \left| 1 + \sum_{i=1}^n \mathbb{I}(z_i = h \ \& \ z_i \neq \max\{\mathcal{L}_{\mathbf{x}}\}), \alpha + \sum_{i=1}^n \mathbb{I}(z_i > h) \right.\right)$, $h = 1, \dots, N_{\max}$;
8. $\pi(\alpha | \dots) = \text{ga}(\alpha | a_{\alpha} + n_{clus} - \mathbb{I}(u > \{O/(1+O)\}), b_{\alpha} - \log(\eta))$, given draws $\eta \sim \text{Beta}(\alpha + 1, n)$, $u \sim \text{Uniform}(0, 1)$, and $O = (a_{\alpha} + n_{clus} - 1) / (\{b_{\alpha} - \log(\eta)\}n)$, where n_{clus} is the number of unique z_i , over $(i = 1, \dots, n)$ (Escobar & West, 1995, p.584).

Standard MCMC Gibbs sampling methods can be used to sample the full conditionals in Steps 1, 2, 7, and 8. The full conditionals in Steps 3 through 6 are each sampled using an adaptive random-walk Metropolis-Hastings algorithm (Roberts & Rosenthal, 2009). The above 8-step sampling algorithm is repeated a large number S of times to construct a discrete-time Harris ergodic Markov chain $\left\{\boldsymbol{\zeta}^{(s)} = (\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{v}, \alpha, \boldsymbol{\gamma}, \boldsymbol{\psi})^{(s)}\right\}_{s=1}^S$, having a posterior distribution $\Pi(\boldsymbol{\zeta}|\mathcal{D}_n)$ as its stationary distribution, provided that a proper prior is assigned to $\boldsymbol{\zeta}$. We have written MATLAB (2012, The MathWorks, Natick, MA) code that implements the described MCMC sampling algorithm.

APPENDIX B: List of Verbal Aggression Items

For the Verbal Aggression Questionnaire, the 24 items are as follows, with each item scored on a 0-2 rating scale.

1. A bus fails to stop for me. I would want to **curse**.
2. A bus fails to stop for me. I would want to **scold**.
3. A bus fails to stop for me. I would want to **shout**.
4. I miss a train because a clerk gave me faulty information. I would want to **curse**.
5. I miss a train because a clerk gave me faulty information. I would want to **scold**.
6. I miss a train because a clerk gave me faulty I would want to **shout**.
7. The grocery store closes just as I am about to enter. I would want to **curse**.
8. The grocery store closes just as I am about to enter. I would want to **scold**.
9. The grocery store closes just as I am about to enter. I would want to **shout**.
10. The operator disconnects me when I had used up my last 10 cents for a call. I would want to **curse**.
11. The operator disconnects me when I had used up my last 10 cents for a call. I would want to **scold**.
12. The operator disconnects me when I had used up my last 10 cents for a call. I would want to **shout**.
13. A bus fails to stop for me. I would **curse**.
14. A bus fails to stop for me. I would **shout**.
15. A bus fails to stop for me. I would **scold**.
16. I miss a train because a clerk gave me faulty information. I would **curse**.
17. I miss a train because a clerk gave me faulty information. I would **scold**.
18. I miss a train because a clerk gave me faulty information. I would **shout**.
19. The grocery store closes just as I am about to enter. I would **curse**.
20. The grocery store closes just as I am about to enter. I would **scold**.
21. The grocery store closes just as I am about to enter. I would **shout**.
22. The operator disconnects me when I had used up my last 10 cents for a call. I would **curse**.

23. The operator disconnects me when I had used up my last 10 cents for a call. I would **scold**.
24. The operator disconnects me when I had used up my last 10 cents for a call. I would **shout**.

Each of the 24 items above is indicated as either a **curse**, **scold**, or **shout** item. Also, items 1-6 and items 13-18 provide the **Other-to-Blame items**. Items 7-12 and items 19-24 provide the **Self to Blame items**. Additionally, items 1-12 are **Behavior Mode: Want** items and items 13-24 are **Behavior Mode: Do** items.

Item	τ_1		τ_2		Modes τ_1	Modes τ_2
	Mean	SD	Mean	SD		
1	-.42	1.27	-.03	1.87	-.05, -.91, 1.40	.56, -.54
2	.06	.83	.20	.85	.05	.22
3	.28	.85	1.09	1.00	.43	1.36
4	-.68	1.47	.09	1.55	-.30, -.90	.56, -1.65, -2.31, -.50, 1.88
5	-.10	.25	.25	.26	-.16	.19
6	.33	1.74	.67	1.21	-.19	.56
7	-.14	.87	1.11	1.43	-.40	1.51
8	.82	.29	2.01	.42	.84	2.04
9	1.52	.52	2.75	.70	1.60	2.78, 3.82
10	-.63	.52	.70	.56	-.77	.67
11	.63	.47	1.29	.59	.71	1.36
12	1.28	1.05	1.70	1.16	1.64	2.02
13	-.61	.46	.21	.47	-.63	.24
14	.14	.72	.63	1.2	-.06	.84
15	1.15	.86	1.69	1.58	1.38, .22	2.23
16	-.25	.92	.20	1.24	-.46	.33
17	.48	.77	1.04	1.35	.46	1.29
18	1.62	1.00	2.17	1.2	1.94	2.47
19	.89	.64	2.12	.93	1.02	2.25
20	.96	.38	2.24	.56	1.10	2.21
21	2.87	.52	3.31	.77	2.92	3.22
22	-.22	.86	.80	1.34	-.48	1.06
23	.61	1.83	1.01	1.27	-.15, 2.67	.64, 1.05
24	2.06	.07	2.56	.89	2.17	2.45

Table 1: For the DDP-RM, the posterior estimates of the ordered category threshold parameters, by item. For the posterior mean and SD estimates, the 95 percent MCCI half-width typically ranged between .00 to .03, with maximum .05.

Model (\underline{m})	$D_1(\underline{m})$	GF(\underline{m})	Pen(\underline{m})
DDP-RM	4984	2008	2976
DP-PCM	5033	2077	2956
3-Mixture PCM	5163	2485	2679
PCM	5716	2783	2934
GPCM	5686	2774	2912
RSM	5726	2791	2936
NRM	5689	2774	2915
GRM	5709	2783	2925

Table 2: The overall mean-squared predictive error, the goodness of fit, and penalty, by model.

Figure Captions

Figure 1. Traceplots of the MCMC posterior samples of the threshold estimates for three items.

Figure 2. Traceplots of the MCMC posterior samples of the ability estimates for six examinees.

Figure 3. The posterior predictive density of the rating category thresholds for three items.

Figure 4. Median, interquartile, and 95-percentile range of the posterior distribution for the neighborhood location (γ) and size (ψ) by item.







